Biometrics 74, 910–923
September 2018

DOI: 10.1111/biom.12859

An Alternative Robust Estimator of Average Treatment Effect in Causal Inference

Jianxuan Liu,[1]* Yanyuan Ma[2,] and Lan Wang[3,]*****

[1]Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, Ohio 43403, U.S.A.
[2]Department of Statistics, Penn State University, University Park, Pennsylvania 16802, U.S.A.
[3]School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455, U.S.A.
email: jianxl@bgsu.edu
**email:* yzm63@psu.edu
***email:* wangx346@umn.edu

Summary. The problem of estimating the average treatment effects is important when evaluating the effectiveness of medical treatments or social intervention policies. Most of the existing methods for estimating the average treatment effect rely on some parametric assumptions about the propensity score model or the outcome regression model one way or the other. In reality, both models are prone to misspecification, which can have undue influence on the estimated average treatment effect. We propose an alternative robust approach to estimating the average treatment effect based on observational data in the challenging situation when neither a plausible parametric outcome model nor a reliable parametric propensity score model is available. Our estimator can be considered as a robust extension of the popular class of propensity score weighted estimators. This approach has the advantage of being robust, flexible, data adaptive, and it can handle many covariates simultaneously. Adopting a dimension reduction approach, we estimate the propensity score weights semiparametrically by using a nonparametric link function to relate the treatment assignment indicator to a low-dimensional structure of the covariates which are formed typically by several linear combinations of the covariates. We develop a class of consistent estimators for the average treatment effect and study their theoretical properties. We demonstrate the robust performance of the estimators on simulated data and a real data example of investigating the effect of maternal smoking on babies' birth weight.

Key words: Average treatment effects; Causal inference; Dimension reduction; Efficient estimators; Propensity score; Robust estimation.

1. Introduction

Estimating the average treatment effect is important in comparing different medical treatments, social programs, and intervention policies. The problem is challenging when the data come from an observational study instead of a randomized experiment. Direct differencing of the sample averages is susceptible to confounding bias, which is caused by imbalances in baseline covariate distributions between the treatment group and the control group.

Under the commonly imposed no unmeasured confounder assumption (Rosenbaum and Rubin, 1983; De Luna et al., 2011), a variety of methods have been proposed to consistently estimate the average treatment effect. The class of doubly robust (DR) estimators (Scharfstein et al., 1999; Robins and Rotnitzky, 2001; Bang and Robins, 2005; Rubin and van der Laan, 2008; Cao et al., 2009; Tan, 2010; van der Laan and Rose, 2011; Rotnitzky et al., 2012; Vansteelandt et al., 2012; van der Laan, 2015; Benkeser and van der Laan, 2016; among others) have been particularly popular due to their double protection against model misspecification.

For most practitioners, the application of DR estimation often adopts parametric specification of both the propensity score model and the outcome regression model, hereafter referred to as parametric DR. The parametric DR estimators are consistent when either the parametric propensity score model or the parametric outcome regressions model is correctly specified. However, Carpenter et al. (2006), Kang and Schafer (2007), and Vansteelandt et al. (2012) observed that the finite-sample bias can be amplified when one of the working models is misspecified and the bias of parametric DR estimators can be severe if both models are slightly misspecified. Vermeulen and Vansteelandt (2015) recently proposed a novel generic strategy for bias reduction under misspecification of both models. Vermeulen and Vansteelandt (2016) further explored the use of data-adaptive estimators in constructing bias-reduced doubly robust estimation. These estimators provide very useful improvement over standard parametric DR estimators, but still need at least one working model to be correctly specified using a parametric model.

Motivated by the practical concern of bias reduction, we propose an alternative approach by directly considering estimators of average treatment effects that are consistent in a larger class of semiparametric propensity score models. The semiparametric class we study imposes a semiparametric structure for the propensity score model while imposing no structure for the outcome regression model. As a direct consequence, our proposed estimator is expected to be consistent for many distributions where most of the standard

parametric DR estimators would become inconsistent. This will be demonstrated by the numerical results in Section 4. Furthermore, we derive the asymptotic normality of the proposed estimator for the average treatment effect, which remains valid for this general class of semiparametric distributions.

There has been growing recent interest in relaxing the parametric specification of working models in parametric DR. Hirano et al. (2003) and Wang et al. (2010) considered non-parametric approach for estimating the propensity score. However, the non-parametric approach is not feasible when many covariates are present due to the curse of dimensionality. Imai and Ratkovic (2014) introduced covariate balancing propensity score as a method that is robust to mild misspecification of the parametric propensity score model. McCaffrey et al. (2004), Ridgeway and McCaffrey (2007), Petersen et al. (2007), Westreich et al. (2010), Lee et al. (2010) explored machine learning approaches for modeling the propensity score but have not studied the asymptotic properties of the resulted average treatment effect estimator. In a sequence of impressive work, van der Laan and his coauthors proposed and carefully studied targeted maximum likelihood estimators (TMLE) which incorporates the state-of-art of machine learning and uses an ensemble of models. See van der Laan and Rubin (2006), the recent manuscript of van der Laan and Rose (2011) and the references therein. van der Laan (2014) showed that a double targeting can guarantee that the bias of the estimator of the target parameter is of second order and hence asymptotically linear. van der Laan (2015) further proposed a general one-step targeted minimum loss-based estimator based on an initial estimator of the nuisance parameters defined by a loss-based super-learner and proved that this one-step TMLE is asymptotically efficient. The latter estimator is understandably more computationally intensive than our proposed approach as it involves multiple tuning parameters and requires cross-validation.

The approach we propose can be viewed as a middle ground between the parametric DR and the non-parametric DR. Compared to parametric DR, our method does not rely on parametric specification of the propensity score model or the outcome regression model. In fact, we do not attempt to model the outcome at all, and only model the propensity score semiparametrically, hence it is more robust as far as the dependence on the propensity score model is concerned. Compared to the non-parametric DR, it has the advantage of being able to handle many covariates. Specifically, we relax the commonly imposed parametric assumption on the propensity score model by only assuming the probability of assigning the treatment depends on the p-dimensional covariate vector $\mathbf{X}$ through several linear combinations $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$, where $\boldsymbol{\beta}$ is a $p \times d$ matrix with $d < p$. We then estimate this conditional probability by employing a non-parametric link function. Note that much work exists in studying how to model the relation between a binary response and many covariates, see, for example, Pregibon (1980), Koenker and Yoon (2009), Li et al. (2016). The special case of $d = 1$ yields the single index model and is especially well studied (Härdle et al., 2004). As an intermediate model for the propensity score in the treatment effect estimation, our semiparametric approach for estimating the propensity score is most closely related to the sufficient dimension reduction literature (Cook, 1998) and is of independent interest. Existing methods for estimating the dimension reduction space such as sliced inverse regression (SIR) (Li, 1991), sliced average variance estimation (SAVE) (Cook and Weisberg, 1991), directional regression (Li and Wang, 2007), generalized directional regression (Li and Dong, 2009; Dong and Li, 2010) have two limitations in relation to our problem. First, they rely mainly on a linearity condition and/or a constant variance condition, that is, $E(\mathbf{X} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$ being a linear function of $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$ and $\mathrm{var}(\mathbf{X} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$ being a constant matrix, or their generalized form, which may not hold in our problem. Second, they require a reversal of the relation between $\mathbf{X}$ and T, that is, they require to compute expectations of the functions of the covariates $\mathbf{X}$ conditional on T. Because T only has two values, each expectation will generate only two different values, which is not sufficient for subsequent operations of these methods. This hampers the direct application of these methods. On the other hand, other methods based on non-parametric regression (Xia, 2007) and semiparametric regression (Ma and Zhu, 2012, 2013) exist, but they also need to be adapted instead of directly applied to estimating the propensity score which concerns binary response.

The rest of the article is organized as follows. In Section 2, we introduce the multi-index semiparametric estimator of the propensity score function and a robust estimator of the average treatment effect. In Section 3, we study the asymptotic properties of the estimators. Simulation studies are conducted and presented in Section 4. We illustrate the usefulness of the method in a real data example of analyzing effect of maternal smoking on babies' birth weight in Section 5 and conclude the article with a brief discussion in Section 6. The Appendix contains the derivation of the efficient score function and the proof of Theorem 1. The regularity conditions, proofs of Lemmas, and additional numerical results are given in the online supplementary document.

2. A Robust Estimator of the Average Treatment Effect

2.1. *Notation and Setup*

We consider the popular setting of a binary treatment T ($T = 1$ for treatment and 0 for control). To estimate the treatment effect, we adopt the potential or counterfactual outcome framework (Neyman et al., 1990; Rubin, 1974). Let $Y^*(1)$ be the outcome of the subject had s/he (possibly counter to fact) received treatment; and $Y^*(0)$ be the outcome of the subject has s/he (possibly counter to fact) received non-treatment. Our goal is to estimate the average treatment effect

$$\tau = E\{Y^*(1) - Y^*(0)\}.$$

The difficulty of the problem arises because for each individual in the sample, we observe either $Y^*(1)$ or $Y^*(0)$, but not both. The observed outcome is $Y = TY^*(1) + Y^*(0)(1 - T)$, that is, the observed outcome is the potential outcome corresponding to the treatment the subject actually receives, which is often referred to as the consistency assumption in causal inference (Rubin, 1986).

Given data from an observational study $\{Y_i, T_i, X_i\}$, $i = 1, \ldots, n$, where Y_i is the response of the ith subject, T_i is the binary treatment indicator, X_i is a vector of covariates, we are interested in estimating the average causal effect of the treatment. Direct differencing the sample averages of the treatment and control groups often leads to a biased estimator of τ in observational studies as the two groups often differ in some covariates that are associated with both the treatment and outcome. Let $\pi(X) = P(T = 1|X)$ be the propensity score function and assume that the unconfoundedness given X assumption is satisfied, that is $\{Y^*(1), Y^*(0)\} \perp T|X$, or the treatment assignment is independent of the potential outcomes given the covariates. Rosenbaum and Rubin (1983) showed that adjusting for propensity score can completely remove the confounding bias from the difference in covariates.

Hahn (1998) derived the semiparametric efficiency bound for estimating τ in the general model where only the independence between treatment and potential outcomes given covariates is assumed. The propensity score can be used in different ways to obtain a consistent estimator for the average treatment effect. Hahn (1998) also proposed an estimator that achieves the semiparametric efficiency bound, but his estimator involves estimating $E(YT|X)$, $E\{Y(1-T)|X\}$ and $\pi(X)$. Hirano et al. (2003) further showed that a simpler estimator that only estimates $\pi(X)$ nonparametrically can also achieve the semiparametric efficiency bound. However, these non-parametric estimators suffer from the curse of dimensionality in real data analysis even with a moderate amount of covariates such as four covariates.

In practice, existing work on causal inference usually adopts a parametric approach to modeling the propensity score function. For example, logistic models are frequently used to model disease occurrence in case-control studies (Prentice and Pyke, 1979; Chatterjee and Carroll, 2005; Lin and Zeng, 2009; Ma and Carroll, 2016), in missing probability problem (Rubin, 1976; Rubin and Little, 2002), and even in survival models (Efron, 1988). However, the parametric approach is prone to model misspecification and can result in substantial bias.

The crux of our robust estimator of the average treatment effect is to develop a flexible estimator of the propensity score function. Instead of the parametric logistic regression model for the propensity score function, we assume

$$\mathrm{pr}(T = 1|\mathbf{X} = \mathbf{x}) = \frac{\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\}}{1 + \exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\}}, \tag{2.1}$$

where $\mathbf{X} \in \mathcal{R}^p$, $\boldsymbol{\beta} \in \mathcal{R}^{p\times d}$, and η is an arbitrary unspecified function. Note that we use the logit link function here for parameterization purpose to ensure that the depicted probability function takes values between 0 and 1. As the function η is completely unspecified, our model allows the probability of being assigned to the treatment to depend on several linear combinations of X in a non-parametric fashion. In contrast, the popular logistic regression model assumes this probability to depend on one particular linear combination of X in a known parametric fashion.

2.2. *Flexible Estimation of the Propensity Score*

To obtain a more concise form, we rewrite (2.1) equivalently as

$$\mathrm{pr}(T = t \mid \mathbf{X} = \mathbf{x}) = \frac{\exp\{t\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\}}{1 + \exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\}}. \tag{2.2}$$

The log-likelihood function of $\boldsymbol{\beta}$ and η is $\sum_{i=1}^{n}(t_i\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i) - \log[1 + \exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\}])$. For identifiability of $\boldsymbol{\beta}$, we require $\boldsymbol{\beta}$ to have the form $\boldsymbol{\beta} = (\mathbf{I}_d, \boldsymbol{\beta}_l^{\mathrm{T}})^{\mathrm{T}}$, where the upper submatrix $\mathbf{I}_d$ is the $d \times d$ identity matrix while the lower submatrix $\boldsymbol{\beta}_l$ is an arbitrary $(p - d) \times d$ matrix. To estimate the semiparametric propensity score function, we need to estimate $\boldsymbol{\beta}_l$ and the unknown function η, the former of which contains $p_t = (p - d)d$ free parameters while the latter can be viewed as an infinite dimensional parameter, where the subindex in p_t stands for "total." In the sequel, for notational convenience, for any arbitrary $p \times d$ matrix $\boldsymbol{\beta}$, we define the concatenation of the columns contained in the lower $p - d$ rows of $\boldsymbol{\beta}$ as $\mathrm{vecl}(\boldsymbol{\beta}) = \mathrm{vec}(\boldsymbol{\beta}_l) = (\beta_{d+1,1}, \ldots, \beta_{p,1}, \ldots, \beta_{d+1,d}, \ldots, \beta_{p,d})^{\mathrm{T}}$ where "vec" stands for vectorization while "vecl" is the vectorization of the lower part of the original matrix.

Our approach of estimation relies on first deriving the influence function using the geometric technique (Bickel et al., 1993; Tsiatis, 2006). In the Appendix, we derive the efficient score function with respect to $\boldsymbol{\beta}$:

$$\mathbf{S}_{\mathrm{eff}}(t_i, \mathbf{x}_i, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i, \eta, \boldsymbol{\eta}') = \mathrm{vecl}\left(\{\mathbf{x}_i - E(\mathbf{X}_i \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\} \times \left[t_i - \frac{\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\}}{1 + \exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\}}\right] \boldsymbol{\eta}'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)^{\mathrm{T}}\right). \tag{2.3}$$

We use the Nadaraya–Watson kernel estimator to estimate $E(\mathbf{X}_i \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)$, that is,

$$\widehat{E}(\mathbf{X} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) = \frac{\sum_{i=1}^{n} \mathbf{x}_i K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})}{\sum_{i=1}^{n} K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})}, \tag{2.4}$$

where h is a bandwidth and K is a multivariate kernel function, $K_h(\cdot) = K(\cdot/h)/h^d$. Neither η nor $\boldsymbol{\eta}'$ is known in a real data analysis. To deal with this complexity, in the following we borrow the idea of locally efficient and adaptively efficient estimators in general and especially in Ma and Zhu (2012) and consider two different options, which lead to two different estimators of $\boldsymbol{\beta}$.

First, we consider an estimator of $\boldsymbol{\beta}$ based on a posited form of η, denoted as η^*, which may not be identical to η. The corresponding derivative is denoted by $\boldsymbol{\eta}^{*\prime}$. This yields the locally efficient score function

$$\mathbf{S}^*_{\mathrm{eff}}(t_i, \mathbf{x}_i, \boldsymbol{\beta}, \eta^*, \boldsymbol{\eta}^{*\prime}) = \mathrm{vecl}\left(\{\mathbf{x}_i - E(\mathbf{X}_i \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\} \times \left[t_i - \frac{\exp\{\eta^*(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\}}{1 + \exp\{\eta^*(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\}}\right] \boldsymbol{\eta}^{*\prime}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)^{\mathrm{T}}\right). \tag{2.5}$$

Obviously, there are many different choices of η, as long as η^* is a smooth function of $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$. For example, when we choose $\eta^*(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) = \mathbf{1}_d^{\mathrm{T}}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ where $\mathbf{1}_d$ is a length d vector of ones. Then $\boldsymbol{\eta}^{*\prime}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) = \mathbf{1}_d$. The locally efficient estimator of $\boldsymbol{\beta}$ solves the estimating equation

$$\sum_{i=1}^{n} \mathrm{vec}\left[\{\mathbf{x}_i - \widehat{E}(\mathbf{X}_i \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\}\left\{t_i - \frac{\exp(\mathbf{1}_d^{\mathrm{T}}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)}{1+\exp(\mathbf{1}_d^{\mathrm{T}}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)}\right\}\mathbf{1}_d^{\mathrm{T}}\right] = \mathbf{0}. \tag{2.6}$$

We denote this estimator by $\widehat{\boldsymbol{\beta}}_1$.

Next, we consider estimating $\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)$ and its first derivative non-parametrically to obtain an adaptively efficient estimator of $\boldsymbol{\beta}$. We adopt the local linear kernel method to estimate $\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ and its first derivative, which solves

$$\sum_{i=1}^{n}\left[t_i - \frac{\exp\{b_0+\mathbf{b}_1^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i-\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0)\}}{1+\exp\{b_0+\mathbf{b}_1^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i-\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0)\}}\right]K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0) = 0 \tag{2.7}$$

$$\sum_{i=1}^{n}\left[t_i - \frac{\exp\{b_0 + \mathbf{b}_1^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0)\}}{1 + \exp\{b_0 + \mathbf{b}_1^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0)\}}\right] \times(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0)K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0) = \mathbf{0}. \tag{2.8}$$

The estimators $\widehat{b}_0$ and $\widehat{\mathbf{b}}_1$ are the estimators of η and $\boldsymbol{\eta}'$ at $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0$, respectively. We can vary $\mathbf{x}_0$ to obtain estimates of the functions at various values. We write the resulting estimators as $\widehat{\eta}(\cdot, \boldsymbol{\beta})$ and $\widehat{\boldsymbol{\eta}}'(\cdot, \boldsymbol{\beta})$, which can be considered as profiled estimators for η and $\boldsymbol{\eta}'$. We subsequently plug $\widehat{\eta}(\cdot, \boldsymbol{\beta})$, $\widehat{\boldsymbol{\eta}}'(\cdot, \boldsymbol{\beta})$, $\widehat{E}(\mathbf{X} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ into (2.3) and solve for $\boldsymbol{\beta}$ to obtain the efficient estimator, which we denote by $\widehat{\boldsymbol{\beta}}_2$.

2.3. *Robust Estimation of the Average Treatment Effect*

To estimate the average treatment effect robustly, we propose to use

$$\widehat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_iY_i}{\widehat{\pi}(X_i)} - \frac{(1-T_i)Y_i}{1-\widehat{\pi}(X_i)}\right\}, \tag{2.9}$$

where $\widehat{\pi}(X_i)$ is obtained from the semiparametric model (2.1) and estimated using either of the two options discussed in Section 2.2. Algorithm 1 below depicts the detailed steps of obtaining the estimator $\widehat{\tau}$ when the locally efficient estimator of $\pi(X_i)$ is used (i.e., based on $\widehat{\boldsymbol{\beta}}_1$). The algorithm based on $\widehat{\boldsymbol{\beta}}_2$ is similar. The above procedure can be considered as an extension of the celebrated Horvitz–Thompson inverse probability weighted estimator (Horvitz and Thompson, 1952), which was originally developed for survey sampling.

The proposed estimator enjoys nice robustness properties. It is more flexible than the parametric propensity score model and hence is less prone to misspecification. It also does not propose any outcome regression models, which leads to computational simplicity. One can further pursue a double robust estimator by augmenting the estimator we propose. It could further improve estimation efficiency at the price of more complex modeling and/or computation. The estimator can accommodate a large number of covariates. Note that although non-parametric smoothing is used to estimate $\pi(X_i)$, the smoothing is implemented with respect to $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$. Under the dimension reduction assumption, it is often sufficient to consider a small d in practice; our estimator does not face the kind of curse of dimensionality that prevents the practical implementation of the estimators in Hahn (1998) and Hirano et al. (2003). Furthermore, we allow the covariate $\mathbf{X}$ to include both continuous and discrete or categorical variables without imposing any distributional assumptions on the covariate.

Algorithm 1 Robust estimator of the average treatment effect

Input: $\{Y_i, T_i, X_i\}$, $i = 1, \ldots, n$, where Y_i is the response of the ith subject, T_i is a binary treatment indicator ($T_i = 1$ for treatment and 0 for control), X_i is a vector of covariates.

Output: Estimator $\widehat{\tau}$.

1: Use (2.6) to obtain a local efficient estimator of $\boldsymbol{\beta}$, denoted as $\widetilde{\boldsymbol{\beta}}$ via, for example, choosing $\eta^*(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) = \mathbf{1}_d^{\mathrm{T}}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$.

2: Perform non-parametric estimation of $\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)$ and its first derivative $\boldsymbol{\eta}'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)$ by implementing (2.7). Write the resulting estimator as $\widehat{\eta}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i, \boldsymbol{\beta})$ and $\widehat{\boldsymbol{\eta}}'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i, \boldsymbol{\beta})$.

3: Perform non-parametric estimation of $E(\mathbf{X}_i \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)$. Write the resulting estimator as $\widehat{E}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)$.

4: Plug $\widehat{\eta}(\cdot, \boldsymbol{\beta})$, $\widehat{\boldsymbol{\eta}}'(\cdot, \boldsymbol{\beta})$ and $\widehat{E}(\cdot)$ in to $\mathbf{S}_{\mathrm{eff}}$ and solve the estimating equation

$$\sum_{i=1}^{n}\mathbf{S}_{\mathrm{eff}}(y_i, \mathbf{x}_i, \boldsymbol{\beta}, \widehat{\eta}, \widehat{\boldsymbol{\eta}}', \widehat{E}) = \mathbf{0},$$

using $\widetilde{\boldsymbol{\beta}}$ as starting value, to obtain the efficient estimator $\widehat{\boldsymbol{\beta}}$.

5: Repeat Step 2 to obtain the final estimator of $\eta(\cdot)$ and form $\widehat{\pi}(X_i) = 1 - 1/[1 + \exp\{\widehat{\eta}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x})\}]$.

6: **return** $\widehat{\tau} = n^{-1}\sum_{i=1}^{n}\left\{\frac{T_iY_i}{\widehat{\pi}(X_i)} - \frac{(1-T_i)Y_i}{1-\widehat{\pi}(X_i)}\right\}$.

REMARK 1. *A technical detail involved in the non-parametric step of the above procedure is bandwidth selection. Through extensive numerical experimentation, we find that the $\boldsymbol{\beta}$ estimation procedure is quite insensitive to the bandwidth, while inference precision could be affected by the bandwidth. Thus, guided by the theoretical properties (see Lemma 1, Lemma 2, and the regularity conditions C4), we recommend simply setting the bandwidth to be $\mathrm{var}(\|X_i\|_2)n^{-1/5}$ throughout the estimation of $\boldsymbol{\beta}$, and use a leave-one-out cross-validation procedure to obtain the smoothing parameter h in estimating η after fixing $\widehat{\boldsymbol{\beta}}$. The same bandwidth then can be used in the inference procedure.*

3. Asymptotic Properties

We now study the asymptotic properties of the estimators for the propensity score function for the robust estimator of the average treatment effect. The regularity conditions that are needed for the theoretical development are given in the online Supplementary Materials. Condition C1 consists of some

standard requirements on the univariate and multivariate kernel functions. Condition C4 contains some mild requirement on the bandwidth. Conditions C2–C3 and C5–C8 contain the boundedness, smoothness, and invertibility of several functions or matrices. All these conditions are very mild.

First, we study the asymptotic properties of $\widehat{\boldsymbol{\beta}}_1$, the nonparametric estimators of $\eta, \boldsymbol{\eta}'$, and $\widehat{\boldsymbol{\beta}}_2$ discussed in Section 2.2. The results are summarized in Lemmas 1 and 2 below. The proofs are relegated to the online Supplementary Materials.

LEMMA 1. *Let $\widehat{\boldsymbol{\beta}}_1$ be the estimator defined in Section 2.2. Under the regularity conditions (C1)–(C6), $\widehat{\boldsymbol{\beta}}_1$ is locally efficient. As $n \to \infty$,*

$$\sqrt{n}\{\mathrm{vecl}(\widehat{\boldsymbol{\beta}}_1) - \mathrm{vecl}(\boldsymbol{\beta})\} \to N\{\mathbf{0}, \mathbf{A}^{-1}\mathbf{G}(\mathbf{A}^{-1})^{\mathrm{T}}\}$$

in distribution, where

$$\mathbf{A} = E\left\{\frac{\partial}{\partial(\mathrm{vec}\,\boldsymbol{\beta})^{\mathrm{T}}}\mathrm{vec}\left(\{\mathbf{X}_i - E(\mathbf{X}_i \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}\left[T_i - \frac{\exp\{\eta^*(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}}{1+\exp\{\eta^*(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}}\right]\right.\right.$$
$$\left.\left.\boldsymbol{\eta}^{*\prime}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)^{\mathrm{T}}\right)\right\},$$

$$\mathbf{G} = E\left\{\mathrm{vecl}\left(\left\{\mathbf{X}_i - E(\mathbf{X}_i \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\right\}\left[T_i - \frac{\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}}{1+\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}}\right]\right.\right.$$
$$\left.\left.\boldsymbol{\eta}'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)^{\mathrm{T}}\right)^{\otimes 2}\right\}.$$

Here and throughout the article, $\mathbf{a}^{\otimes 2} \equiv \mathbf{a}\mathbf{a}^{\mathrm{T}}$.

LEMMA 2. *Assume the regularity conditions (C1)–(C4) and (C7)–(C8) hold. The local linear kernel estimators of $\widehat{\eta}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ and $\widehat{\boldsymbol{\eta}}'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ defined in Section 2.2 satisfy*

$$E\{\widehat{\eta}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\} - \eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) = O(h^m), \quad E\{\widehat{\boldsymbol{\eta}}'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\} - \boldsymbol{\eta}'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) = O(h^m),$$
$$\mathrm{var}\{\widehat{\eta}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\} = O_p\{(nh^d)^{-1}\}, \quad \mathrm{var}\{\widehat{\boldsymbol{\eta}}'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\} = O_p\{(nh^{d+2})^{-1}\}.$$

Furthermore, $\widehat{\boldsymbol{\beta}}_2$ defined in Section 2.2 is efficient and satisfies

$$\sqrt{n}\{\mathrm{vecl}(\widehat{\boldsymbol{\beta}}_2) - \mathrm{vecl}(\boldsymbol{\beta}_2)\} \to N(\mathbf{0}, \mathbf{V}^{-1})$$

in distribution as $n \to \infty$, where

$$\mathbf{V} = E\{\mathbf{S}_{\mathrm{eff}}(T_i, \mathbf{X}_i, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \eta, \boldsymbol{\eta}', E)^{\otimes 2}\}$$
$$= E\left\{\mathrm{vecl}\left(\left\{\mathbf{X}_i - E(\mathbf{X}_i \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\right\}\right.\right.$$
$$\left.\left.\times\left[T_i - \frac{\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}}{1+\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}}\right]\boldsymbol{\eta}'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)^{\mathrm{T}}\right)^{\otimes 2}\right\}.$$

We provide the asymptotic property of the average treatment estimator $\widehat{\tau}$ defined in Section 2.3, where the propensity is based on the dimension reduction estimation. We adopt two standard assumptions in causal inference, that is, no unmeasured confounding and positivity.

THEOREM 1. *Under the regularity conditions (C1)–(C8), when $n \to \infty$ the estimator $\widehat{\tau}$ from (2.9) based on $\widehat{\boldsymbol{\beta}}_2$ satisfies*

$$\sqrt{n}(\widehat{\tau} - \tau) \to N(0, \sigma^2)$$

in distribution, where $\sigma^2 = \sigma^2_{\mathrm{eff}} + \mathbf{a}^{\mathrm{T}}E(\mathbf{S}_{\mathrm{eff}}\mathbf{S}_{\mathrm{eff}}^{\mathrm{T}})^{-1}\mathbf{a}$, with

$$\sigma^2_{\mathrm{eff}} = \mathrm{var}\left[\left\{\frac{T_iY_i}{\pi(\mathbf{X}_i)} - \frac{(1-T_i)Y_i}{1-\pi(\mathbf{X}_i)} - \tau\right\}\right.$$
$$\left. - \left\{\frac{Y_i^*(1)}{\pi(\mathbf{X}_i)} + \frac{Y_i^*(0)}{1-\pi(\mathbf{X}_i)}\right\}\{T_i - \pi(\mathbf{X}_i)\}\right],$$
$$\mathbf{a} = E\big(\left[Y_{i1}\{1-\pi(\mathbf{X}_i)\} + Y_i^*(0)\pi(\mathbf{X}_i)\right]\boldsymbol{\eta}'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) \otimes \mathbf{X}_{iL}\big).$$

REMARK 2. *In the above asymptotic variance expression, σ^2_{eff} is the optimal estimation variance (Hahn, 1998; Hirano et al., 2003) when the propensity is completely unknown and estimated purely non-parametrically. The additional term is the price we pay when we use a dimension reduction procedure to estimate π instead of doing it fully non-parametrically. In other words, our estimator is in general not efficient.*

Regarding the theoretical efficiency bound in estimating a treatment effect, whether the propensity score is completely known or completely unknown, the efficiency bound in estimating the average treatment effect is the same as is given in Hahn (1998). In our context, the propensity score is partially known, in that we know it has the dimension reduction structure. Thus, the efficiency bound in estimating the treatment effect should be in between the completely known and completely unknown cases, and hence is also the same as that given in Hahn (1998).

Regarding achieving optimal efficiency, if inverse probability weighting method is used, the efficiency bound is only achieved if the propensity score is estimated non-parametrically, regardless of whether the true propensity score is known or not known (Hirano et al., 2003). Thus, in the setting where the propensity score is partially known, the efficiency bound is still only achieved if the estimator ignores the fact that the propensity score is partially known and is estimated non-parametrically. If instead the known knowledge about the propensity is used in the estimator, then the asymptotic efficiency is strictly larger than the efficiency bound.

However, estimating the propensity score non-parametrically is often infeasible in practice, especially when there are many covariates. Thus, a natural compromise is to adopt as flexible a model as possible, such as the dimension reduction model, to facilitate the propensity score estimation, which provides a trade-off between efficiency and practical applicability.

4. Monte Carlo Studies

4.1. *A Simulation Study on Estimating the Propensity Score Function*

We first conduct a simulation study to investigate the performance of the flexible semiparametric estimators proposed in Section 2.2 for the propensity score.

We generate the vector of covariates $\mathbf{X} = (X_1, X_2, X_3, X_4, X_5, X_6)^T$ as follows. The covariates X_1 and X_2 are generated from independent standard normal distributions. We let $X_3 = 0.2X_1 + 0.2(X_2 + 2.0)^2$, $X_4 = 0.1 + 0.2(X_1 + X_2) + 0.3(X_1 + 1)^2$, and generate X_5 and X_6 independently from Bernoulli distribution with success probability $\exp(X_3)/\{1 + \exp(X_3)\}$ and $\exp(X_4)/\{1 + \exp(X_4)\}$, respectively. In (2.2), we consider the following two different functional forms:

- Setting (1): $\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) = \sin(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$, where $d = 1$ and $\boldsymbol{\beta} = (1.0, -1.2, 0.8, -1.7, -1.5, 0.5)^{\mathrm{T}}$.
- Setting (2): $\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) = \sin(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}) + \sin(\boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{x})$, where $d = 2$, $\boldsymbol{\beta}_1 = (1.0, 0.0, 1.2, 0.8, -1.2, 0.8)^{\mathrm{T}}$ and $\boldsymbol{\beta}_2 = (0.0, 1.0, 1.3, 0.7, 1.1, -0.7)^{\mathrm{T}}$.

For comparison purposes, we implement the oracle estimator and compare with our proposed semiparametric estimators $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$. The oracle estimator assumes the functional form of η in (2.2) is known, although $E(\mathbf{x}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ is still estimated through the kernel regression in (2.4). Even though the oracle estimator is unrealistic, it provides a benchmark since it is the best performance one could expect to obtain. The local estimator $\widehat{\boldsymbol{\beta}}_1$ replaces η with a mis-specified function in the estimation procedure and estimates $E(\mathbf{x}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ non-parametrically. We posit the models $\eta^*(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) = \sin(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + 0.8) - 0.3$ and $\eta^*(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) = \sin(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x} + 0.5) + \cos(\boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{x} - 0.5)$ for setting (1) and (2), respectively. The efficient estimator $\widehat{\boldsymbol{\beta}}_2$ does not use any posited model for η. Instead, we estimate $E(\mathbf{x}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$, η and $\boldsymbol{\eta}'$ through non-parametric regression, that is, we followed the algorithm described in Section 2. The efficient estimator $\widehat{\boldsymbol{\beta}}_2$ is more computationally involved since it solves estimating equations to obtain the non-parametric components η and $\boldsymbol{\eta}'$ at n locations inside the search for $\boldsymbol{\beta}$ which does not have a closed form. To alleviate the computational burden, we performed the non-parametric estimation on a set of grid points and then performed a linear interpolation for $d = 1$ and a bilinear interpolation for $d = 2$ to obtain the values at each $\widehat{\boldsymbol{\beta}}^{(k)\mathrm{T}}\mathbf{x}_i$, where $\widehat{\boldsymbol{\beta}}^{(k)}$ represents the kth iteration of the estimated $\widehat{\boldsymbol{\beta}}$ during solving the estimating equation in Step 4 of the algorithm in Section 2.

We repeat each experiment 1000 times with sample size $n = 500$ and 1000, respectively. The results are summarized in Table 1 for setting (1) and Table 2 for setting (2). From Table 1, we observe that the performance of both $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$ is close to that of the oracle estimator. All estimators have very small bias for both sample sizes. The results in the table also provide the median of the estimated standard errors using

Table 1

The median, the sample standard errors (std), the average of the estimated standard deviation ($\widehat{std}$) and the coverage of the estimated 95% confidence interval (CI) of the oracle estimator, locally efficient estimator and the efficient estimator, of $\boldsymbol{\beta}$ in simulated setting (1).

		$\boldsymbol{\beta}_2$	$\boldsymbol{\beta}_3$	$\boldsymbol{\beta}_4$	$\boldsymbol{\beta}_5$	$\boldsymbol{\beta}_6$
	True	−1.2	0.8	−1.7	−1.5	0.5
Oracle	Median	−1.2000	0.7760	−1.6932	−1.5000	0.4964
$n = 500$	$\widehat{\text{std}}$	0.3044	0.3885	0.2019	0.3854	0.3117
	std	0.3406	0.4116	0.2300	0.3800	0.3670
	CI	0.9320	0.9230	0.9220	0.9610	0.9630
Local	Median	−1.0224	0.6503	−1.7137	−1.4016	0.4694
$n = 500$	$\widehat{\text{std}}$	0.2897	0.3431	0.2411	0.5357	0.3581
	std	0.3726	0.4450	0.3736	0.5194	0.4415
	CI	0.8680	0.8830	0.8660	0.9440	0.9300
Efficient	Median	−1.2155	0.8105	−1.6986	−1.5037	0.5070
$n = 500$	$\widehat{\text{std}}$	0.5674	0.7080	0.3036	0.5353	0.4337
	std	0.4735	0.4813	0.4129	0.5211	0.5074
	CI	0.9750	0.9860	0.8850	0.9540	0.9440
Oracle	Median	−1.1879	0.8133	−1.6843	−1.5061	0.5000
$n = 1000$	$\widehat{\text{std}}$	0.2106	0.2444	0.1435	0.2684	0.2097
	std	0.2405	0.2906	0.1506	0.2924	0.2234
	CI	0.9400	0.9310	0.9400	0.9510	0.9640
Local	Median	−1.1802	0.7920	−1.6926	−1.3853	0.4710
$n = 1000$	$\widehat{\text{std}}$	0.2369	0.2463	0.1419	0.2748	0.2196
	std	0.2720	0.2755	0.1874	0.2931	0.2698
	CI	0.9240	0.9430	0.9430	0.9210	0.9450
Efficient	Median	−1.1936	0.8030	−1.6999	−1.4953	0.4966
$n = 1000$	$\widehat{\text{std}}$	0.3963	0.3656	0.1712	0.3716	0.2364
	std	0.2561	0.2337	0.1724	0.3168	0.2165
	CI	0.9590	0.9720	0.9400	0.9320	0.9520

Table 2

The median, the sample standard errors (std), the median of the estimated standard deviation ($\widehat{std}$) and the coverage of the estimated 95% confidence interval (CI) of the oracle estimator, locally efficient estimator and the efficient estimator of $\boldsymbol{\beta}$ in simulated setting (2).

		$\boldsymbol{\beta}_{13}$	$\boldsymbol{\beta}_{14}$	$\boldsymbol{\beta}_{15}$	$\boldsymbol{\beta}_{16}$	$\boldsymbol{\beta}_{23}$	$\boldsymbol{\beta}_{24}$	$\boldsymbol{\beta}_{25}$	$\boldsymbol{\beta}_{26}$
	True	1.2	0.8	−1.2	0.8	1.3	0.7	1.1	−0.7
Oracle	Median	1.1874	0.8112	−1.1817	0.8318	1.3251	0.7152	1.0779	−0.7113
$n = 500$	$\widehat{\text{std}}$	0.2085	0.2057	0.3807	0.3622	0.2215	0.2251	0.3949	0.3704
	std	0.2703	0.2861	0.4262	0.4070	0.2873	0.2871	0.4411	0.4085
	CI	0.9380	0.9260	0.9570	0.9610	0.9290	0.9230	0.9690	0.9580
Local	Median	1.1939	0.7960	−1.1061	0.8663	1.3070	0.6214	1.2427	−0.7372
$n = 500$	$\widehat{\text{std}}$	0.3259	0.3271	0.5747	0.5526	0.4377	0.4376	0.8213	0.7297
	std	0.3440	0.3748	0.5479	0.5553	0.5138	0.4981	0.7111	0.6871
	CI	0.9270	0.9210	0.9610	0.9700	0.9110	0.9670	0.9490	0.9390
Efficient	Median	1.2292	0.8759	−1.2214	0.8315	1.3566	0.7002	1.0998	−0.7723
$n = 500$	$\widehat{\text{std}}$	0.7113	0.6808	0.6997	0.7027	0.6757	0.5555	0.6938	0.6622
	std	0.6104	0.4356	0.4836	0.4129	0.5529	0.5078	0.5195	0.4764
	CI	0.9200	0.9700	0.9770	0.9930	0.9540	0.9260	0.9630	0.9690
Oracle	Median	1.1928	0.8154	−1.2070	0.8194	1.3053	0.7098	1.0877	−0.6931
$n = 1000$	$\widehat{\text{std}}$	0.1460	0.1423	0.2620	0.2458	0.1437	0.1447	0.2629	0.2435
	std	0.1742	0.1710	0.2852	0.2700	0.1690	0.1647	0.2778	0.2591
	CI	0.9610	0.9480	0.9560	0.9610	0.9420	0.9540	0.9540	0.9650
Local	Median	1.2028	0.7792	−1.0987	0.8109	1.3363	0.6012	1.3007	−0.7161
$n = 1000$	$\widehat{\text{std}}$	0.2224	0.1970	0.3610	0.3381	0.2816	0.2865	0.6493	0.5385
	std	0.2551	0.2402	0.4031	0.3784	0.2226	0.3547	0.5111	0.4654
	CI	0.9450	0.9440	0.9470	0.9550	0.9490	0.9670	0.9620	0.9550
Efficient	Median	1.2208	0.8606	−1.2053	0.8055	1.3637	0.7079	1.0827	−0.7109
$n = 1000$	$\widehat{\text{std}}$	0.4734	0.4541	0.3217	0.3032	0.4532	0.3743	0.4572	0.3026
	std	0.4529	0.3142	0.3351	0.2927	0.4261	0.2521	0.3789	0.2716
	CI	0.9230	0.9730	0.9380	0.9450	0.9310	0.9780	0.9570	0.9660

the results in Lemma 1 and Lemma 2 and the empirical coverage probability of the 95% confidence intervals. These results indicate that the asymptotic normal approximation is accurate for the sample sizes. We observe similar performance in Table 2. The standard errors of the $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$ become smaller as the sample size grows and the confidence interval coverage probabilities become closer to the nominal level.

4.2. *Additional Simulations*

We further compare the performance of estimators for higher dimensional covariates where $p = 10$. In dimension reduction literature, $p = 10$ is considered to be rather high dimension. See Ma and Zhu (2012) for an investigation of covariate dimension issues. We independently generate X_1, X_2 from Uniform$(0, 1)$, X_3 and X_7 from Normal$(0, 0.5^2)$, X_4 from Normal$(0, 1)$. Then we form $X_5 = X_1 + X_2X_4$, $X_6 = -X_2 + X_1X_3$, $X_8 = (X_4 - X_2)X_1 - X_7$, $X_9 = X_1X_7 - X_3$ and $X_{10} = X_2X_8 - X_9$. Further, we explore the situation which the covariate dimension cannot be much reduced. We set the true propensity score function to be $\text{pr}(T = 1 \mid \mathbf{X}) = 1 - [1 + \exp\{0.1\sqrt{\mid X_5 - X_4^2 \mid} - \cos(X_2X_8) - X_1\exp(X_6) - (X_3 - X_9)X_7 + \exp(-X_{10})\}]^{-1}$ and the true outcome function to be $Y = -T\exp(-X_{10}) + \sin(X_1 + X_2) - X_3^2 - \cos(X_4) - X_5 + X_7\log(X_6^2) - \cos(X_8) - T \mid X_9 \mid$. We now examine the performance of various method in terms of estimating the average treatment effect. To implement our semiparametric estimator, we set $d = 1$, which is certainly not the case in the true model, and investigated the locally efficient estimation procedure, where we posit a mis-specified model $\eta^*(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) = 0.4\cos(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$. This $\eta^*(\cdot)$ restricts the function value to $[-0.4, 0.4]$ while the true value is out of this range. We summarize the estimated average treatment effect in Figure 1, where "true" represents the result when the true propensity score is used, "η^*" is the from the semiparametric estimator when "$\eta^*(\cdot)$" is used in the local estimation of $\boldsymbol{\beta}$, "$\widehat{\eta}$" is when the link function η is estimated as described in the efficient estimator procedure. We also compare our semiparametric approach with targeted maximum likelihood estimation (TMLE) (van der Laan and Rubin, 2006), the biased reduced double robust (BRdr) estimator proposed by Vermeulen and Vansteelandt (2015), Tan's improved method (Tan) (Tan, 2006, 2010), and the standard method where the propensity score is estimated via logistic regression (Logistic). From the data generating process described above, it is clear that neither the propensity score model nor the outcome model is a generalized linear model. In implementing the TMLE method, rather than providing a model for either the propensity score or the outcome, we

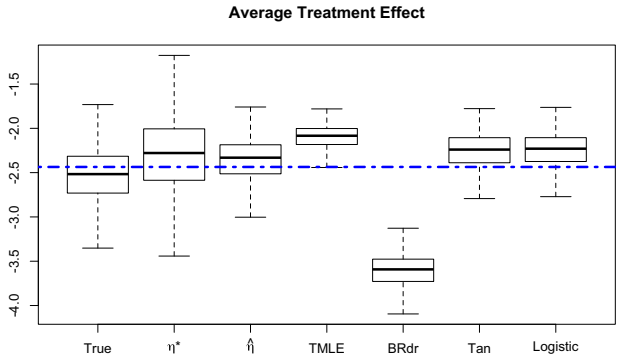

Figure 1. Average treatment effect when $p = 10$, no dimension reduction is possible and the outcome is $Y = -T\exp(-X_{10}) + \sin(X_1 + X_2) - X_3^2 - \cos(X_4) - X_5 + X_7 \log(X_6^2) - \cos(X_8) - T \mid X_9 \mid$. Dimension reduction model is used by setting $d = 1$. The dashed line is the true average treatment effect.

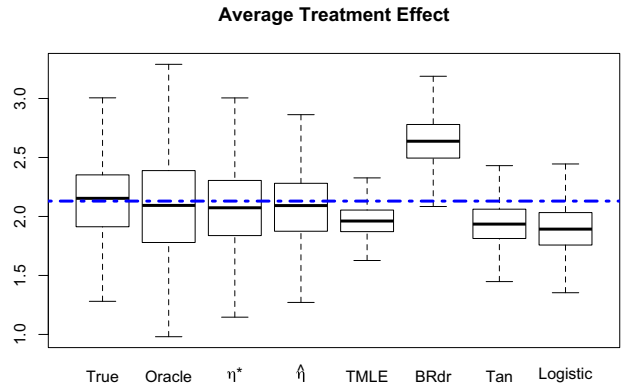

Figure 2. Average treatment effect when $p = 10$, dimension reduction structure valid for $d = 1$ and the outcome is $Y = \exp(T + X_{10}) + \sin(X_1)X_2 + X_3^2 - \cos(X_4 - X_5) + \log(X_6^2)X_7 + X_8 - TX_9$. The dashed line is the true average treatment effect.

allow the TMLE algorithm to call the powerful SuperLearner to estimate these two quantities in a data adaptive fashion. Our implementation of BRdr estimator is also data adaptive, please see Vermeulen and Vansteelandt (2015) for detail. To estimate the propensity scores, a logistic regression model is assumed according to Tan's description of the method. In contrast, Tan's method uses generalized linar model (glm) in estimating the treatment outcome, hence the outcome model is misspecified since the true model here is not a glm. From Figure 1, although the TMLE method has the smallest variance, it is severely biased. BRdr, Tan, and Logistic are also biased.

We further examine the situation when the covariate dimension indeed can be reduced to $d = 1$. Specifically, we set $\boldsymbol{\beta} = (1.0, -0.4, 0.4, -0.2, -0.2, 0.4, 0.3, -0.3, -0.6, -0.6)^{\mathrm{T}}$ and the true η function $\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) = \exp(0.5\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\cos(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ to generate the treatment T, where $\mathbf{x}$ is generated in the same way as before. We generate the outcome from the model $Y = \exp(T + X_{10}) + \sin(X_1)X_2 + X_3^2 - \cos(X_4 - X_5) + \log(X_6^2)X_7 + X_8 - TX_9$. We implemented the same estimators as before, and added an oracle estimator where the true η is used in estimating $\boldsymbol{\beta}$. From the results in Figure 2, we see that the semiparametric methods still outperform other estimators.

4.3. *Comparison of Several Methods in Data by Kang–Schafer*

Finally, we examine the efficient and locally efficient estimator on the data generated following Kang and Schafer (2007). Specifically, we generated $(Z_1, Z_2, Z_3, Z_4)^{\mathrm{T}}$ from Normal$(\mathbf{0}, \mathbf{I}_4)$ and then form $x_1 = \exp(z_1/2), x_2 = z_2/\{1 + \exp(z_1)\}, x_3 = (z_1z_3/25 + 0.6)^3, x_4 = (z_2 + z_4 + 20)^2/400$. The outcome model is generated as $y = 210 + 27.4z_1 + 13.7z_2 + 13.7z_3 + 13.7z_4 + \epsilon$, where $\epsilon \sim N(0, 1)$ and the true propensity function is $\pi = \mathrm{expit}(-z_1 + 0.5z_2 - 0.25z_3 - 0.1z_4)$. We use the observable data (Y_i, T_i, X_i) for $i = 1, 2, \ldots, n$ to estimate the propensity score $\widehat{\pi}_i$ for $i = 1, 2, \ldots, n$, then calculate the average treatment effect $\widehat{\tau}$. The performance of the average treatment effect can be found in Figure 3, where "True" refers to the average treatment effect calculated from an inverse probability weighted method where the true weight is used. Both the locally efficient and efficient estimators yield reasonable results in comparison with other methods, regardless of whether $d = 1$ or $d = 2$.

5. A Real Data Example

We next apply the proposed semiparametric methods to analyze the average effect of maternal smoking on babies' birth weight. The Low Birth Weight data constitute observations from mothers in Pennsylvania, USA and contain birth information on 4642 infants (Cattaneo, 2010). This dataset was originally used by Almond et al. (2005). The outcome of interest Y is infant birth weight measured in grams. The binary variable T denotes the mother's smoking status (1 = smoking, 0 = non-smoking). The covariates include mother's age, mother's marital status, an indicator variable for alcohol consumption during pregnancy, an indicator for whether there was a previous birth where the newborn died, mother's education, father's education, number of prenatal

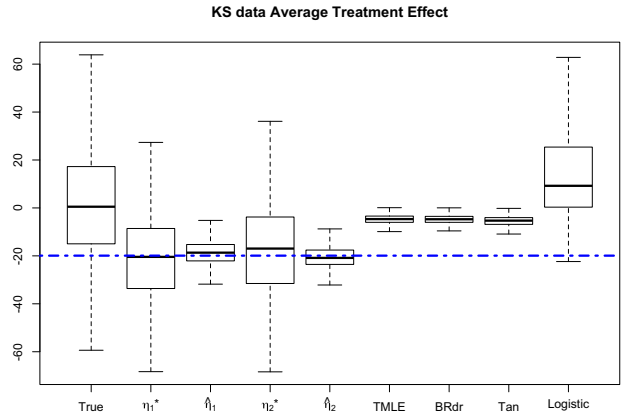

Figure 3. Average treatment effect on Kang and Schafer data. η_1^* and $\widehat{\eta}_1$ are for $d = 1$. η_2^* and $\widehat{\eta}_2$ are for $d = 2$. The dashed line is the true average treatment effect.

Table 3
Low Birth Weight data example

	Naive			Local			Efficient		
	Est	Std	P-value	Est	Std	P-value	Est	Std	P-value
(Intercept)	0.9848	0.2631	0.0002						
Age	1.0021	0.3607	0.0055						
mmarried	−0.9480	0.1030	0.0000	−0.8020	0.2187	0.0002	−1.6922	0.3537	0.0000
alcohol	1.5886	0.1844	0.0000	1.5021	0.4156	0.0003	2.7014	0.4683	0.0000
deadkids	0.3893	0.0909	0.0000	0.4070	0.1232	0.0010	0.4980	0.1554	0.0014
medu	−0.0964	0.0190	0.0000	−0.0675	0.0281	0.0164	−0.2066	0.0571	0.0003
fedu	−0.0426	0.0118	0.0003	−0.0499	0.0182	0.0061	−0.1067	0.0369	0.0038
nprenatal	−0.0299	0.0111	0.0071	−0.0346	0.0141	0.0143	−0.0513	0.0221	0.0204
monthslb	0.0062	0.0015	0.0000	0.0062	0.0019	0.0012	0.0097	0.0028	0.0007
mrace	0.6888	0.1184	0.0000	0.7607	0.2093	0.0003	1.1446	0.2421	0.0000
fbaby	−0.2574	0.1059	0.0150	−0.2728	0.1181	0.0209	−0.3799	0.1881	0.0435

Table 4
Average treatment difference in the Low Birth Weight data. Bootstrap mean (BS mean) and Bootstrap std (BS std). Bootstrap sample $B = 1000$

	Naive	Efficient	Local	Logistic	TMLE	BRdr	Tan
Estimate	−275.25	−295.77	−306.32	−352.08	−219.96	−228.89	−230.57
BS mean	−275.10	−292.85	−304.69	−352.11	−219.69	−229.33	−231.34
BS std	21.36	38.62	54.50	46.78	29.50	29.34	27.66

care visits, mother's race, indicator of first born baby, and months since last birth (monthslb).

Based on data from the 4642 infants, the naive average weight difference of the two groups of babies belonging to smoking and non-smoking mothers yields −275.25 grams. Considering that this naive result is not necessarily a valid estimator of the causal result of smoking on birth weight, we next studied the proposed estimators. Specifically, we compare three estimators of average treatment effect discussed in the Section 4: "Logistic," $\hat{\tau}_1$, and $\hat{\tau}_2$. The estimated propensity score functions are summarized in Table 3. The estimated average treatment difference corresponding to "Logistic," $\hat{\tau}_1$, and $\hat{\tau}_2$ are −352.08, −295.77, and −306.32 grams, respectively. In addition, we compare the average causal effect with Tan's improved method, TMLE and BRdr. The results

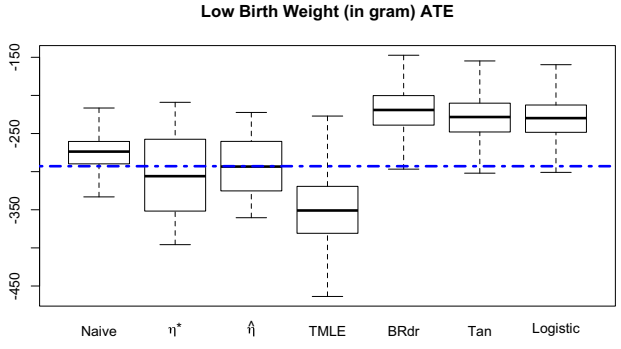

Figure 4. Bootstrap average treatment effect. The dashed line is the mean of the average treatment effect calculated from the efficient estimation procedure.

indicate that maternal smoking has a negative impact on babies' birth weight. The estimate average treatment differences are summarized in Table 4 along with the mean and standard deviation from 1000 bootstrap samples for each method. The bootstrap average treatment effect from the seven approaches can be found in Figure 4. Note that the estimator using propensity score estimated by logistic regression is substantially different from $\hat{\tau}_1$ and $\hat{\tau}_2$. This suggests logistic regression may not provide an adequate model for the propensity score function.

6. Conclusion and Discussions

In this article, we propose a semiparametric approach to estimating the average treatment effect. The approach is less prone to propensity score model misspecification compared to the logistic regression-based inverse probability weighted estimators, which have dominant roles in causal inference. A parametric propensity score model (e.g., logistic regression model) is certainly a lot more informative than a semiparametric model such as the dimension reduction model we propose, but it also bears a greater risk of being misspecified. If the parametric propensity score model is misspecified, then the resulting estimation of the average treatment effect is inconsistent. Furthermore, the semiparametric estimator does not rely on specification of the outcome regression model, and hence is attractive when a reliable outcome regression model is hard to obtain and/or compute, such as when studying treatment effects on complex diseases. We note that if one is willing to propose outcome models and carry more computation, then further extending our method to a doubly robust estimator could bring additional benefit such as efficiency gain.

It is of interest to investigate whether a dimension reduction propensity score model will always lead to more efficient treatment effect estimation than a parametric one, in the case that both models are correct. However, we find that it is not true in general. The relation can go either way, and it depends on the specific models. We summarize the results in Lemma 3 in the online Supplementary Materials.

Not able to find any definitive relation between the dimension reduction model and a general parametric model, we further investigate the situation of nested models. For the sake of comparing two models that are both correct, this certainly makes much sense. To this end, the model will be the same as in (2.2), except that now η is a known function. Unfortunately, even for this case, as shown in the online Supplementary Materials, there is no definitive relation we can claim. Thus, even when the parametric model is a submodel of the dimension reduction model, there is no definitive relation between the two estimators of the average treatment effect based on the two models. Our intuition is that not only the model makes a difference, but also the specific estimator used in the propensity score model has a role to play. The overall picture is unclear and is potentially very complex; much work is needed to fully understand these relations and can lead to interesting research results.

Finally, even though our initial intention is to overcome the potential issue of mis-specification of both the propensity score model and the outcome regression model through employing a more relaxed modeling strategy of the former and giving up modeling of the latter, and subsequently proposing inverse property weighting, double robust estimator can be used in combination with our method to further gain efficiency. As it is well known in the original form of the double robust estimator, in combination with the semiparametric propensity score model, when the treatment response is modeled correctly, the method will be more efficient than our method. If the treatment response is modeled incorrectly, depending on how "wrong" the model is, the method could be less efficient than our method. However, if the method of Tan (2010) is adopted, in combination with the semiparametric propensity score model, one can always obtain a more efficient estimator than our method, regardless of whether the treatment response is modeled correctly or not. Thus, to achieve improved efficiency, one can strive to propose a "good" model for the treatment response, and further perform additional computation to obtain the correlation adjustment required in Tan (2010).

7. Supplementary Materials

Regularity conditions in 3 and proofs of Lemmas referenced in Sections 3 and 6, additional numerical results for Section 4 and the dataset on birth weight mentioned in Section 5 are available with this article on the *Biometrics* website on Wiley Online Library.

Acknowledgements

Yanyuan Ma was partially supported by NSF grant DMS-1608540 and NINDS grant NS073671. Lan Wang was partially supported by NSF grant DMS-1512267 and DMS-1712706.

References

Almond, D., Chay, K. Y., and Lee, D. S. (2005). The costs of low birth weight. *Quarterly Journal of Economics* **120**, 1031–1083.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**, 962–973.

Benkeser, D. and Van Der Laan, M. (2016). The Highly Adaptive Lasso Estimator. In *Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on*, 689–696. IEEE.

Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore, MD: The Johns Hopkins University Press.

Cao, W., Tsiatis, A. A., and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika* **96**, 723–734.

Carpenter, J. R., Kenward, M. G., and Vansteelandt, S. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* **169**, 571–584.

Cattaneo, M. D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics* **155**, 138–154.

Chatterjee, N. and Carroll, R. J. (2005). Semiparametric maximum likelihood estimation in case-control studies of gene-environment interactions. *Biometrika* **92**, 399–418.

Cook, D. R. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*. New York: Wiley.

Cook, D. R. and Weisberg, S. (1991). Discussion of sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86**, 28–33.

De Luna, X., Waernbaum, I., and Richardson, T. (2011). Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika* **98**, 861–875.

Dong, Y. and Li, B. (2010). Dimension reduction for non-elliptically distributed predictors: Second-order moments. *Biometrics* **97**, 279–294.

Efron, B. (1988). Logistic regression, survival analysis, and the kaplan-meier curve. *Journal of the American Statistical Association* **83**, 414–425.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66**, 315–331.

Härdle, W., Werwatz, A., Müller, M., and Sperlich, S. (2004). *Nonparametric and Semiparametric Models*. New York: Springer.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71**, 1161–1189.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association* **47**, 663–685.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **76**, 243–263.

Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* **22**, 523–539.

Koenker, R. and Yoon, J. (2009). Parametric links for binary choice models: A fisherian-bayesian colloquy. *Journal of Econometrics* **152**, 120–130.

Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine* **29**, 337–346.

Li, B. and Dong, Y. (2009). Dimension reduction for non-elliptically distributed predictors. *The Annuals of Statistics* **37**, 1272–1298.

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association* **102**, 997–1008.

Li, D., Wang, X., Lin, L., and Dey, D. K. (2016). Flexible link functions in nonparametric binary regression with gaussian process priors. *Biometrics* **72**, 707–719.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86**, 316–327.

Lin, D. Y. and Zeng, D. (2009). Proper analysis of secondary phenotype data in case-control association studies. *Genetic Epidemiology* **33**, 256–265.

Ma, Y. and Carroll, R. J. (2016). Semiparametric estimation in the secondary analysis of case–control studies. *Journal of the Royal Statistical Society, Series B* **78**, 127–151.

Ma, Y. and Zhu, L. (2012). A semiparametric approach to dimension reduction. *Journal of the American Statistical Association* **107**, 168–179.

Ma, Y. and Zhu, L. (2013). Efficient estimation in sufficient dimension reduction. *The Annuals of Statistics* **41**, 250–268.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods* **9**, 403–425.

Neyman, J., Dabrowska, D. M., and Speed, T. P. (1990). On the application of probability theory to agricultural experiments: Essay on principles, section 9. *Statistical Scince* **5**, 465–480.

Petersen, M. L., Wang, Y., Van Der Laan, M. J., Guzman, D., Riley, E., and Bangsberg, D. R. (2007). Pillbox organizers are associated with improved adherence to hiv antiretroviral therapy and viral suppression: A marginal structural model analysis. *Clinical Infectious Diseases* **45**, 908–915.

Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Journal of the Royal Statistical Society, Series C* **29**, 15–23.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–411.

Ridgeway, G. and McCaffrey, D. F. (2007). Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Sciences* **22**, 540–543.

Robins, J. M. and Rotnitzky, A. (2001). Comment on the bickel and kwon article, inference for semiparametric models: Some questions and an answer. *Statistica Sinica* **11**, 920–936.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Rotnitzky, A., Lei, Q., Sued, M., and Robins, J. M. (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika* **99**, 439–456.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology* **66**, 688–701.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.

Rubin, D. B. (1986). Which ifs have causal answers. *Journal of the American Statistical Association* **81**, 961–962.

Rubin, D. B. and Little, R. A. (2002). *Statistical Analysis with Missing Data (2nd ed.)*. New York: Wiley.

Rubin, D. B. and van der Laan, M. J. (2008). Empirical efficiency maximization: Improved locally efficient covariate adjustment in randomized experiments and survival analysis. *The International Journal of Biostatistics* **4**, Article–5.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* **94**, 1096–1120.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association* **101**, 1619–1637.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika* **97**, 661–682.

Tsiatis, A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.

van der Laan, M. (2015). A generally efficient targeted minimum loss based estimator. U.C. Berkeley Division of Biostatistics Working Paper Series.

van der Laan, M. J. (2014). Targeted estimation of nuisance parameters to obtain valid statistical inference. *The International Journal of Biostatistics* **10**, 29–57.

van der Laan, M. J. and Rose, S. (2011). *Targeted Learning*. New York: Springer.

van der Laan, M. J. and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics* **2**, 1–40.

Vansteelandt, S., Bekaert, M., and Claeskens, G. (2012). On model selection and model misspecification in causal inference. *Statistical Methods in Medical Research* **21**, 7–30.

Vermeulen, K. and Vansteelandt, S. (2015). Bias-reduced doubly robust estimation. *Journal of the American Statistical Association* **110**, 1024–1036.

Vermeulen, K. and Vansteelandt, S. (2016). Data-adaptive bias-reduced doubly robust estimation. *The International Journal of Biostatistics* **12**, 253–282.

Wang, L., Rotnitzky, A., and Lin, X. (2010). Nonparametric regression with missing outcomes using weighted kernel estimating equations. *Journal of the American Statistical Association* **105**, 1135–1146.

Westreich, D., Lessler, J., and Funk, M. J. (2010). Propensity score estimation: Neural networks, support vector machines, decision trees (cart), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology* **63**, 826–833.

Xia, Y. C. (2007). A constructive approach to the estimation of dimension reduction directions. *Annals of Statistics* **35**, 2654–2690.

Received July 2016. *Revised November* 2017.
Accepted December 2017.

Appendix

A.1. *Derivation of the Efficient Score Function*

Taking derivative with respect to $\boldsymbol{\beta}$ of the logarithm of the probability density function function, it is easy to verify that the score function with respect to $\boldsymbol{\beta}$ is

$$\mathbf{S}_{\boldsymbol{\beta}}(T_i, \mathbf{x}_i, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i, \eta, \boldsymbol{\eta}') = \mathrm{vecl}\left(\mathbf{x}_i\left[T_i - \frac{\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\}}{1+\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\}}\right]\boldsymbol{\eta}'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)^{\mathrm{T}}\right).$$

The efficient score is the residual after projecting the score vector with respect to $\boldsymbol{\beta}$ onto the nuisance tangent space Λ

(Tsiatis, 2006). The nuisance tangent space, denoted Λ, is the mean-squared closure of all nuisance tangent spaces of all parametric submodels. We can verify that

$$\Lambda = \left(\left[T - \frac{\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}}{1+\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}} \right] \mathbf{a}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) : \ \forall \mathbf{a}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) \in \mathcal{R}^{(p-d)\times d} \right)$$

We then obtain its orthogonal complement

$$\Lambda^{\perp} = \Big[\mathbf{f}(Y,\mathbf{X}) : \ \forall \mathbf{f} \in \mathcal{R}^{(p-d)d} \ s.t. \ E\{\mathbf{f}(1,\mathbf{X}) \mid T=1, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\} \times \frac{\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}}{1+\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}} = F\{\mathbf{f}(0,\mathbf{X}) \mid T=0, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\} \Big].$$

We now write

$$\begin{aligned}
&\mathbf{S}_{\boldsymbol{\beta}}(T_i, \mathbf{x}_i, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i, \eta, \boldsymbol{\eta}') \\
&= \mathrm{vecl}\left(\mathbf{x}_i \left[T_i - \frac{\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\}}{1+\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\}} \right] \boldsymbol{\eta}'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)^{\mathrm{T}} \right) \\
&= \mathrm{vecl}\left(E(\mathbf{X} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) \left[T_i - \frac{\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\}}{1+\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\}} \right] \boldsymbol{\eta}'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)^{\mathrm{T}} \right) \\
&+ \mathrm{vecl}\left(\mathbf{x} - E(\mathbf{X} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) \left[T_i - \frac{\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\}}{1+\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\}} \right] \boldsymbol{\eta}'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)^{\mathrm{T}} \right).
\end{aligned}$$

We can readily verify that

$$\mathrm{vecl}\left(E(\mathbf{X} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) \left[T_i - \frac{\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\}}{1+\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\}} \right] \boldsymbol{\eta}'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)^{\mathrm{T}} \right) \in \Lambda$$

and

$$\mathrm{vecl}\left(\mathbf{x} - E(\mathbf{X} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) \left[T_i - \frac{\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\}}{1+\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\}} \right] \boldsymbol{\eta}'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)^{\mathrm{T}} \right) \in \Lambda^{\perp},$$

hence this yields the desired result.

A.2. *Proof of Theorem 1*

From (2.9), we write

$$\begin{aligned}
&n^{1/2}(\widehat{\tau} - \tau) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \frac{Y_i T_i}{\pi(\mathbf{X}_i)} - \frac{Y_i(1-T_i)}{1-\pi(\mathbf{X}_i)} - \tau \right\} \\
&\quad + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \frac{Y_i T_i}{\widehat{\pi}(\mathbf{X}_i)} - \frac{Y_i T_i}{\pi(\mathbf{X}_i)} - \frac{Y_i(1-T_i)}{1-\widehat{\pi}(\mathbf{X}_i)} + \frac{Y_i(1-T_i)}{1-\pi(\mathbf{X}_i)} \right\} \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \frac{Y_i T_i}{\pi(\mathbf{X}_i)} - \frac{Y_i(1-T_i)}{1-\pi(\mathbf{X}_i)} - \tau \right\} \\
&\quad + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\frac{Y_i T_i}{\pi^2(\mathbf{X}_i)} \{\pi(\mathbf{X}_i) - \widehat{\pi}(\mathbf{X}_i)\} \right. \\
&\quad \left. - \frac{Y_i(1-T_i)}{\{1-\pi(\mathbf{X}_i)\}^2} \{\widehat{\pi}(\mathbf{X}_i) - \pi(\mathbf{X}_i)\} \right] \\
&\quad + O_p\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{\widehat{\pi}(\mathbf{X}_i) - \pi(\mathbf{X}_i)\}^2 \right] \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \frac{Y_i T_i}{\pi(\mathbf{X}_i)} - \frac{Y_i(1-T_i)}{1-\pi(\mathbf{X}_i)} - \tau \right\} \\
&\quad - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\frac{Y_i T_i}{\pi^2(\mathbf{X}_i)} + \frac{Y_i(1-T_i)}{\{1-\pi(\mathbf{X}_i)\}^2} \right] \{\widehat{\pi}(\mathbf{X}_i) - \pi(\mathbf{X}_i)\} \\
&\quad + O_p(n^{1/2}h^{2m} + n^{-1/2}h^{-d}).
\end{aligned}$$

Now

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\frac{Y_i T_i}{\pi^2(\mathbf{X}_i)} + \frac{Y_i(1-T_i)}{\{1-\pi(\mathbf{X}_i)\}^2} \right] \{\widehat{\pi}(\mathbf{X}_i) - \pi(\mathbf{X}_i)\} \\
&\quad = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\frac{Y_i T_i}{\pi^2(\mathbf{X}_i)} + \frac{Y_i(1-T_i)}{\{1-\pi(\mathbf{X}_i)\}^2} \right] \\
&\qquad \times \left[\frac{\exp\{\widehat{\eta}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i)\}}{1+\exp\{\widehat{\eta}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i)\}} - \frac{\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}}{1+\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}} \right] \\
&\quad = T_1 + T_2 + T_2,
\end{aligned}$$

where

$$\begin{aligned}
T_1 &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\frac{Y_i T_i}{\pi^2(\mathbf{X}_i)} + \frac{Y_i(1-T_i)}{\{1-\pi(\mathbf{X}_i)\}^2} \right] \\
&\qquad \times \left[\frac{\exp\{\eta(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i)\}}{1+\exp\{\eta(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i)\}} - \frac{\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}}{1+\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}} \right], \\
T_2 &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\frac{Y_i T_i}{\pi^2(\mathbf{X}_i)} + \frac{Y_i(1-T_i)}{\{1-\pi(\mathbf{X}_i)\}^2} \right] \\
&\qquad \times \left[\frac{\exp\{\widehat{\eta}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}}{1+\exp\{\widehat{\eta}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}} - \frac{\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}}{1+\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}} \right], \\
T_3 &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\frac{Y_i T_i}{\pi^2(\mathbf{X}_i)} + \frac{Y_i(1-T_i)}{\{1-\pi(\mathbf{X}_i)\}^2} \right] \\
&\qquad \times \left[\frac{\exp\{\widehat{\eta}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i)\}}{1+\exp\{\widehat{\eta}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i)\}} - \frac{\exp\{\eta(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i)\}}{1+\exp\{\eta(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i)\}} \right] \\
&\qquad - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\frac{Y_i T_i}{\pi^2(\mathbf{X}_i)} + \frac{Y_i(1-T_i)}{\{1-\pi(\mathbf{X}_i)\}^2} \right] \\
&\qquad \times \left[\frac{\exp\{\widehat{\eta}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}}{1+\exp\{\widehat{\eta}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}} - \frac{\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}}{1+\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}} \right].
\end{aligned}$$

It is easy to see that

$$\begin{aligned}
T_3 &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\frac{Y_i T_i}{\pi^2(\mathbf{X}_i)} + \frac{Y_i(1-T_i)}{\{1-\pi(\mathbf{X}_i)\}^2} \right] \\
&\qquad \times \frac{\partial}{\partial \boldsymbol{\beta}} \left[\frac{\exp\{\widehat{\eta}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}}{1+\exp\{\widehat{\eta}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}} - \frac{\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}}{1+\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}} \right] \Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})
\end{aligned}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_iT_i}{\pi^2(\mathbf{X}_i)}+\frac{Y_i(1-T_i)}{\{1-\pi(\mathbf{X}_i)\}^2}\right]\left(\frac{\exp\{\widehat{\eta}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}\widehat{\eta}'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)}{[1+\exp\{\widehat{\eta}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}]^2}\right.$$

$$\left.-\frac{\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}\eta'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)}{[1+\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}]^2}\right)^{\mathrm{T}}\Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}\otimes\mathbf{X}_{iL}^{\mathrm{T}}\sqrt{n}\mathrm{vecl}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})$$

$$= o_p(1),$$

where the last equality is because $\sqrt{n}\mathrm{vecl}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) = O_p(1)$ based on Lemma 2, and because of the consistency of $\widehat{\eta}, \widehat{\eta}'$ established in Lemma 2.

It is also easy to see that

$$T_1 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_iT_i}{\pi^2(\mathbf{X}_i)}+\frac{Y_i(1-T_i)}{\{1-\pi(\mathbf{X}_i)\}^2}\right]$$

$$\times\frac{\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}\eta'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)}{[1+\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}]^2}^{\mathrm{T}}\otimes\mathbf{X}_{iL}^{\mathrm{T}}\sqrt{n}\mathrm{vecl}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})+o_p(1)$$

$$= E\left(\left[\frac{Y_iT_i}{\pi^2(\mathbf{X}_i)}+\frac{Y_i(1-T_i)}{\{1-\pi(\mathbf{X}_i)\}^2}\right]\pi(\mathbf{X}_i)\{1-\pi(\mathbf{X}_i)\}\eta'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)^{\mathrm{T}}\otimes\mathbf{X}_{iL}^{\mathrm{T}}\right)$$

$$\times\frac{1}{\sqrt{n}}\sum_{i=1}^{n}E(\mathbf{S}_{\mathrm{eff}}\mathbf{S}_{\mathrm{eff}})^{-1}\mathbf{S}_{\mathrm{eff}}(\mathbf{X}_i,T_i)+o_p(1)$$

$$= E\left(\left[Y_i^*(1)\{1-\pi(\mathbf{X}_i)\}+Y_i^*(0)\pi(\mathbf{X}_i)\right]\eta'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)^{\mathrm{T}}\otimes\mathbf{X}_{iL}^{\mathrm{T}}\right)$$

$$\times\frac{1}{\sqrt{n}}\sum_{i=1}^{n}E(\mathbf{S}_{\mathrm{eff}}\mathbf{S}_{\mathrm{eff}})^{-1}\mathbf{S}_{\mathrm{eff}}(\mathbf{X}_i,T_i)+o_p(1)$$

$$= \mathbf{a}^{\mathrm{T}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}E(\mathbf{S}_{\mathrm{eff}}\mathbf{S}_{\mathrm{eff}}^{\mathrm{T}})^{-1}\mathbf{S}_{\mathrm{eff}}(\mathbf{X}_i,T_i)+o_p(1),$$

where $Y_i^*(1)$ and $Y_i^*(0)$ are potential outcomes under treatment and no treatment respectively, and we used the independence assumption between potential outcomes and treatment in the second last equality.

We now analyze T_2. To this end, with the same notation as in the proof of Lemma 2 in the online supplement,

$$T_2 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{Y_iT_i}{\pi^2(\mathbf{X}_i)}+\frac{Y_i(1-T_i)}{\{1-\pi(\mathbf{X}_i)\}^2}\right]$$

$$\times\left[\frac{\exp\{\widehat{\eta}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}}{1+\exp\{\widehat{\eta}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}}-\frac{\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}}{1+\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}}\right]$$

$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{Y_i^*(1)}{\pi(\mathbf{X}_i)}+\frac{Y_i^*(0)}{1-\pi(\mathbf{X}_i)}\right\}\left\{\widehat{H}(\mathbf{t}_i)-H(\mathbf{t}_i)\right\}.$$

Here again we used the independence assumption in the last equality. We consider $\widehat{H}(\mathbf{t}_i)$ as the direct kernel estimator of $H(\mathbf{t}_i)$, that is, $\widehat{H}(\mathbf{t}_i) = \{\sum_{j=1}^{n}K_h(\mathbf{t}_j-\mathbf{t}_i)Y_j\}/\{\sum_{j=1}^{n}K_h(\mathbf{t}_j-\mathbf{t}_i)\}$. We further obtain

$$T_2 = \frac{1}{n^{3/2}}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\frac{Y_i^*(1)}{H(\mathbf{t}_i)}+\frac{Y_i^*(0)}{1-H(\mathbf{t}_i)}\right\}\left\{\frac{K_h(\mathbf{t}_j-\mathbf{t}_i)Y_j}{n^{-1}\sum_{k=1}^{n}K_h(\mathbf{t}_k-\mathbf{t}_i)}-H(\mathbf{t}_i)\right\}$$

$$= \frac{1}{n^{3/2}}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\frac{Y_i^*(1)}{H(\mathbf{t}_i)}+\frac{Y_i^*(0)}{1-H(\mathbf{t}_i)}\right\}$$

$$\times\left[\frac{K_h(\mathbf{t}_j-\mathbf{t}_i)Y_j}{f(\mathbf{t}_i)}\left\{1-\frac{n^{-1}\sum_{k=1}^{n}K_h(\mathbf{t}_k-\mathbf{t}_i)-f(\mathbf{t}_i)}{f(\mathbf{t}_i)}\right\}-H(\mathbf{t}_i)\right]$$

$$+O_p(n^{1/2}h^{2m}+n^{-1/2}h^{-d})$$

$$= \frac{1}{n^{3/2}}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\frac{Y_i^*(1)}{H(\mathbf{t}_i)}+\frac{Y_i^*(0)}{1-H(\mathbf{t}_i)}\right\}\left\{\frac{K_h(\mathbf{t}_j-\mathbf{t}_i)Y_j}{f(\mathbf{t}_i)}-H(\mathbf{t}_i)\right\}$$

$$-\frac{1}{n^{3/2}}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\frac{Y_i^*(1)}{H(\mathbf{t}_i)}+\frac{Y_i^*(0)}{1-H(\mathbf{t}_i)}\right\}H(\mathbf{t}_i)\left\{\frac{K_h(\mathbf{t}_j-\mathbf{t}_i)-f(\mathbf{t}_i)}{f(\mathbf{t}_i)}\right\}+o_p(1)$$

$$= \frac{1}{n^{3/2}}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\frac{Y_i^*(1)}{H(\mathbf{t}_i)}+\frac{Y_i^*(0)}{1-H(\mathbf{t}_i)}\right\}\left[\frac{K_h(\mathbf{t}_j-\mathbf{t}_i)}{f(\mathbf{t}_i)}\{Y_j-H(\mathbf{t}_i)\}\right]+o_p(1)$$

$$= n^{-1/2}\sum_{i=1}^{n}\left\{\frac{Y_i^*(1)}{H(\mathbf{t}_i)}+\frac{Y_i^*(0)}{1-H(\mathbf{t}_i)}\right\}E\left[\frac{K_h(\mathbf{t}_j-\mathbf{t}_i)}{f(\mathbf{t}_i)}\{Y_j-H(\mathbf{t}_i)\}\,|\,\mathbf{t}_i,T_i\right]$$

$$+n^{-1/2}\sum_{j=1}^{n}E\left(\left\{\frac{Y_i^*(1)}{H(\mathbf{t}_i)}+\frac{Y_i^*(0)}{1-H(\mathbf{t}_i)}\right\}\left[\frac{K_h(\mathbf{t}_j-\mathbf{t}_i)}{f(\mathbf{t}_i)}\{Y_j-H(\mathbf{t}_i)\}\right]|\,\mathbf{t}_j,Y_j\right)$$

$$-n^{1/2}E\left(\left\{\frac{Y_i^*(1)}{H(\mathbf{t}_i)}+\frac{Y_i^*(0)}{1-H(\mathbf{t}_i)}\right\}\left[\frac{K_h(\mathbf{t}_j-\mathbf{t}_i)}{f(\mathbf{t}_i)}\{Y_j-H(\mathbf{t}_i)\}\right]\right)+o_p(1)$$

$$= n^{-1/2}\sum_{j=1}^{n}E\left(\left\{\frac{Y_i^*(1)}{H(\mathbf{t}_i)}+\frac{Y_i^*(0)}{1-H(\mathbf{t}_i)}\right\}\left[\frac{K_h(\mathbf{t}_j-\mathbf{t}_i)}{f(\mathbf{t}_i)}\{Y_j-H(\mathbf{t}_i)\}\right]|\,\mathbf{t}_j,Y_j\right)$$

$$+o_p(1)$$

$$= n^{-1/2}\sum_{i=1}^{n}\left\{\frac{Y_i^*(1)}{H(\mathbf{t}_i)}+\frac{Y_i^*(0)}{1-H(\mathbf{t}_i)}\right\}\{T_i-H(\mathbf{t}_i)\}+o_p(1)$$

$$= n^{-1/2}\sum_{i=1}^{n}\left\{\frac{Y_i^*(1)}{\pi(\mathbf{X}_i)}+\frac{Y_i^*(0)}{1-\pi(\mathbf{X}_i)}\right\}\{T_i-\pi(\mathbf{X}_i)\}+o_p(1).$$

Combining the above results regarding T_1, T_2, and T_3, we obtain

$$n^{1/2}(\widehat{\tau}-\tau) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\left\{\frac{Y_iT_i}{\pi(\mathbf{X}_i)}-\frac{Y_i(1-T_i)}{1-\pi(\mathbf{X}_i)}-\tau\right\}\right.$$

$$\left.-\left\{\frac{Y_i^*(1)}{\pi(\mathbf{X}_i)}+\frac{Y_i^*(0)}{1-\pi(\mathbf{X}_i)}\right\}\{T_i-\pi(\mathbf{X}_i)\}\right]$$

$$-\mathbf{a}^{\mathrm{T}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}E(\mathbf{S}_{\mathrm{eff}}\mathbf{S}_{\mathrm{eff}}^{\mathrm{T}})^{-1}\mathbf{S}_{\mathrm{eff}}(\mathbf{X}_i,T_i)+o_p(1). \tag{A.1}$$

Comparing with the results in Hirano et al. (2003), it is now clear that the component in (A.1) is the efficient influence function, while the remaining component in the expansion of $n^{1/2}(\widehat{\tau}-\tau)$ is the difference between the influence functions of our estimator and the efficient estimator, hence is orthogonal to the efficient influence function. In fact the orthogonality is also easily checked by direct calculation. □

A.3. *Statement of Lemma 3*

LEMMA 3. *Assume the treatment allocation is independent of the potential treatment outcome given the covariates. Assume further that the probability of treatment is bounded away from 0 and 1. Assume a parametric model $\pi(\mathbf{X}_i, \boldsymbol{\gamma})$ with true parameter value $\boldsymbol{\gamma}_0$. Then when $n \to \infty$, the estimator $\widehat{\tau}$ from (2.9) satisfies $\sqrt{n}(\widehat{\tau} - \tau) \to N(0, \sigma^2)$, where $\sigma^2 = \sigma_{\text{eff}}^2 + E(B_i^2)$ where σ_{eff}^2 is the same as in Theorem 1, and $B_i = \left\{ \frac{Y_i^*(1)}{\pi(\mathbf{X}_i)} + \frac{Y_i^*(0)}{1-\pi(\mathbf{X}_i)} \right\} \{T_i - \pi(\mathbf{X}_i)\} - E\left(\left[\frac{Y_i^*(1)}{\pi(\mathbf{X}_i)} + \frac{Y_i^*(0)}{1-\pi(\mathbf{X}_i)} \right] \frac{\partial \pi(\mathbf{X}_i, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}_0} \right) \phi(\mathbf{X}_i, T_i)$, where $\phi(\mathbf{X}_i, T_i)$ is the influence function of $\widehat{\boldsymbol{\gamma}}$.*

A.4. *Comparing Average Treatment Effect Estimators for Nested Propensity Models*

When η is a known function, the efficient score function for $\boldsymbol{\beta}$ is

$$\begin{aligned}\widetilde{\mathbf{S}}_{\text{eff}}(y_i, \mathbf{x}_i, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i) &= \operatorname{vecl}\left(\mathbf{x}_i \left[y_i - \frac{\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\}}{1+\exp\{\eta(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\}} \right] \boldsymbol{\eta}'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)^{\mathrm{T}} \right) \\ &= \operatorname{vecl}\left[\mathbf{x}_i \left\{ y_i - \pi(\mathbf{X}_i) \right\} \boldsymbol{\eta}'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)^{\mathrm{T}} \right] \\ &= \left\{ y_i - \pi(\mathbf{X}_i) \right\} \boldsymbol{\eta}'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i) \otimes \mathbf{x}_{iL},\end{aligned}$$

and the efficient influence function is $E(\widetilde{\mathbf{S}}_{\text{eff}}\widetilde{\mathbf{S}}_{\text{eff}}^{\mathrm{T}})^{-1}\widetilde{\mathbf{S}}_{\text{eff}}$. Using the results in Lemma 3, we have

$$\begin{aligned}B_i &= \left\{ \frac{Y_i^*(1)}{\pi(\mathbf{X}_i)} + \frac{Y_i^*(0)}{1-\pi(\mathbf{X}_i)} \right\} \{T_i - \pi(\mathbf{X}_i)\} \\ &\quad - E\left(\left[\frac{Y_i^*(1)}{\pi(\mathbf{X}_i)} + \frac{Y_i^*(0)}{1-\pi(\mathbf{X}_i)} \right] \pi(\mathbf{X}_i)\{1-\pi(\mathbf{X}_i)\} \boldsymbol{\eta}'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)^{\mathrm{T}} \otimes \mathbf{X}_{iL}^{\mathrm{T}} \right) \\ &\quad \times E(\widetilde{\mathbf{S}}_{\text{eff}}\widetilde{\mathbf{S}}_{\text{eff}}^{\mathrm{T}})^{-1}\widetilde{\mathbf{S}}_{\text{eff}} \\ &= \left\{ \frac{Y_i^*(1)}{\pi(\mathbf{X}_i)} + \frac{Y_i^*(0)}{1-\pi(\mathbf{X}_i)} \right\} \{T_i - \pi(\mathbf{X}_i)\} \\ &\quad - E\left([Y_i^*(1)\{1-\pi(\mathbf{X}_i)\} + Y_i^*(0)\pi(\mathbf{X}_i)] \boldsymbol{\eta}'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)^{\mathrm{T}} \otimes \mathbf{X}_{iL}^{\mathrm{T}} \right) \\ &\quad \times E(\widetilde{\mathbf{S}}_{\text{eff}}\widetilde{\mathbf{S}}_{\text{eff}}^{\mathrm{T}})^{-1}\widetilde{\mathbf{S}}_{\text{eff}} \\ &= \left\{ \frac{Y_i^*(1)}{\pi(\mathbf{X}_i)} + \frac{Y_i^*(0)}{1-\pi(\mathbf{X}_i)} \right\} \{T_i - \pi(\mathbf{X}_i)\} \\ &\quad - \mathbf{a}^{\mathrm{T}} E(\widetilde{\mathbf{S}}_{\text{eff}}\widetilde{\mathbf{S}}_{\text{eff}}^{\mathrm{T}})^{-1} \{\boldsymbol{\eta}'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) \otimes \mathbf{X}_{iL}\} \left\{ T_i - \pi(\mathbf{X}_i) \right\},\end{aligned}$$

Now let

$$C_i \equiv B_i + \mathbf{a}^{\mathrm{T}} E(\mathbf{S}_{\text{eff}}\mathbf{S}_{\text{eff}}^{\mathrm{T}})^{-1}\mathbf{S}_{\text{eff}}(\mathbf{X}_i, T_i)$$

$$\begin{aligned}&= \left\{ \frac{Y_i^*(1)}{\pi(\mathbf{X}_i)} + \frac{Y_i^*(0)}{1-\pi(\mathbf{X}_i)} \right\} \{T_i - \pi(\mathbf{X}_i)\} \\ &\quad - \mathbf{a}^{\mathrm{T}} E(\widetilde{\mathbf{S}}_{\text{eff}}\widetilde{\mathbf{S}}_{\text{eff}}^{\mathrm{T}})^{-1} \{\boldsymbol{\eta}'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) \otimes \mathbf{X}_{iL}\} \left\{ T_i - \pi(\mathbf{X}_i) \right\} \\ &\quad + \mathbf{a}^{\mathrm{T}} E(\mathbf{S}_{\text{eff}}\mathbf{S}_{\text{eff}}^{\mathrm{T}})^{-1} \{\boldsymbol{\eta}'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) \otimes \{\mathbf{X}_{iL} - E(\mathbf{X}_{iL} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\} \left\{ T_i - \pi(\mathbf{X}_i) \right\} \\ &= \left[\left\{ \frac{Y_i^*(1)}{\pi(\mathbf{X}_i)} + \frac{Y_i^*(0)}{1-\pi(\mathbf{X}_i)} \right\} - \mathbf{a}^{\mathrm{T}} E(\widetilde{\mathbf{S}}_{\text{eff}}\widetilde{\mathbf{S}}_{\text{eff}}^{\mathrm{T}})^{-1} \{\boldsymbol{\eta}'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) \otimes \mathbf{X}_{iL}\} \right. \\ &\quad \left. + \mathbf{a}^{\mathrm{T}} E(\mathbf{S}_{\text{eff}}\mathbf{S}_{\text{eff}}^{\mathrm{T}})^{-1} \{\boldsymbol{\eta}'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) \otimes \{\mathbf{X}_{iL} - E(\mathbf{X}_{iL} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\} \right] \left\{ T_i - \pi(\mathbf{X}_i) \right\}.\end{aligned}$$

Now, following the previous notation to let $\mathbf{t}_i = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i$, and $H(\mathbf{t}_i) = \pi(\mathbf{X}_i)$,

$$\begin{aligned}&E\{C_i \mathbf{a}^{\mathrm{T}} E(\mathbf{S}_{\text{eff}}\mathbf{S}_{\text{eff}}^{\mathrm{T}})^{-1}\mathbf{S}_{\text{eff}}(\mathbf{X}_i, T_i)\} \\ &= E\left(\left[\left\{ \frac{Y_i^*(1)}{\pi(\mathbf{X}_i)} + \frac{Y_i^*(0)}{1-\pi(\mathbf{X}_i)} \right\} - \mathbf{a}^{\mathrm{T}} E(\widetilde{\mathbf{S}}_{\text{eff}}\widetilde{\mathbf{S}}_{\text{eff}}^{\mathrm{T}})^{-1} \{\boldsymbol{\eta}'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) \otimes \mathbf{X}_{iL}\} \right. \right. \\ &\quad \left. \left. + \mathbf{a}^{\mathrm{T}} E(\mathbf{S}_{\text{eff}}\mathbf{S}_{\text{eff}}^{\mathrm{T}})^{-1} \{\boldsymbol{\eta}'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) \otimes \{\mathbf{X}_{iL} - E(\mathbf{X}_{iL} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\} \right] \left\{ T_i - \pi(\mathbf{X}_i) \right\}^2 \right. \\ &\quad \left. \times \mathbf{a}^{\mathrm{T}} E(\mathbf{S}_{\text{eff}}\mathbf{S}_{\text{eff}}^{\mathrm{T}})^{-1} \{\boldsymbol{\eta}'(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) \otimes \{\mathbf{X}_{iL} - E(\mathbf{X}_{iL} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\} \right) \\ &= E\left(\left[\left\{ \frac{Y_i^*(1)}{H(\mathbf{t}_i)} + \frac{Y_i^*(0)}{1-H(\mathbf{t}_i)} \right\} - \mathbf{a}^{\mathrm{T}} E(\widetilde{\mathbf{S}}_{\text{eff}}\widetilde{\mathbf{S}}_{\text{eff}}^{\mathrm{T}})^{-1} \{\boldsymbol{\eta}'(\mathbf{t}_i) \otimes \mathbf{X}_{iL}\} \right. \right. \\ &\quad \left. \left. + \mathbf{a}^{\mathrm{T}} E(\mathbf{S}_{\text{eff}}\mathbf{S}_{\text{eff}}^{\mathrm{T}})^{-1} \boldsymbol{\eta}'(\mathbf{t}_i) \otimes \{\mathbf{X}_{iL} - E(\mathbf{X}_{iL} \mid \mathbf{t}_i)\} \right] H(\mathbf{t}_i) \left\{ 1 - H(\mathbf{t}_i) \right\} \right. \\ &\quad \left. \times \mathbf{a}^{\mathrm{T}} E(\mathbf{S}_{\text{eff}}\mathbf{S}_{\text{eff}}^{\mathrm{T}})^{-1} \boldsymbol{\eta}'(\mathbf{t}_i) \otimes \{\mathbf{X}_{iL} - E(\mathbf{X}_{iL} \mid \mathbf{t}_i)\} \right) \\ &= E\left(\left[\left\{ \frac{Y_i^*(1)}{H(\mathbf{t}_i)} + \frac{Y_i^*(0)}{1-H(\mathbf{t}_i)} \right\} \right. \right. \\ &\quad \left. \left. - \mathbf{a}^{\mathrm{T}} \{E(\widetilde{\mathbf{S}}_{\text{eff}}\widetilde{\mathbf{S}}_{\text{eff}}^{\mathrm{T}})^{-1} - E(\mathbf{S}_{\text{eff}}\mathbf{S}_{\text{eff}}^{\mathrm{T}})^{-1}\} \boldsymbol{\eta}'(\mathbf{t}_i) \otimes \{\mathbf{X}_{iL} - E(\mathbf{X}_{iL} \mid \mathbf{t}_i)\} \right] \right. \\ &\quad \left. \times H(\mathbf{t}_i) \left\{ 1 - H(\mathbf{t}_i) \right\} \mathbf{a}^{\mathrm{T}} E(\mathbf{S}_{\text{eff}}\mathbf{S}_{\text{eff}}^{\mathrm{T}})^{-1} \boldsymbol{\eta}'(\mathbf{t}_i) \otimes \{\mathbf{X}_{iL} - E(\mathbf{X}_{iL} \mid \mathbf{t}_i)\} \right),\end{aligned}$$

which is not necessarily zero. Thus, there is no definitive relation we can say even when the parametric model is a submodel of the dimension reduction model.